誤用や話し言葉に対応可能なコーパス分析システムにおけるタグ検索機能

山本裕子（愛知淑徳大学）川村よし子（東京国際大学）
ラニガン・マシュー（LendingHome）小森早江子（中部大学）本間妙（中部大学）

Spoken Language and Error Support in Corpus Analysis Using Tag Search

Hiroko YAMAMOTO, Aichi Shukutoku University
Yoshiko KAWAMURA, Tokyo International University Matthew LANIGAN, LendingHome
Saeko KOMORI, Chubu Universtiy Tae HOMMA, Chubu University

要旨

本研究は、形態素解析における誤解析にも対応可能なコーパス分析システムの開発をめざす。筆者らは、教師自らが集めたデータをコーパスとして分析できるツール、Co-Chuの開発を進めてきた。Co-Chuには【Build】【Import】【Edit】【Analyze】の４機能がある。話し言葉等のデータを取り込み、形態素解析に誤解析が生じた際には【Edit】機能を活用し、データにタグを付与することで適切に形態素解析が行える。今回、上記の【Analyze】機能に「タグ検索機能」を加え、タグを種別ごとに検索できる仕組みを整えた。この機能は、言い誤りや言い淀みの分析、若者ことばや方言使用の実態調査などに活用可能である。本発表では、このタグ検索機能の概要を示すとともに、運用実験の結果について報告する。

キーワード：コーパス分析システムCo-Chu,形態素解析,誤解析,タグ付け,タグ検索機能

１．はじめに

本研究では、形態素解析における誤解析にも対応可能なタグ付けを行うことのできるコーパス分析システムの開発をめざしている。日本語教師や日本語研究者が分析したいテキストは、文法的に正しく、形の整っている文章だけではない。学習者の作文のように誤用の含まれた文章や言い誤りや方言等を含んだ会話文も含まれている。そのため、形態素解析がうまく行かない場合も多い。しかも、既存の学習者コーパスや会話コーパスではなく、独自のコーパスを用いて分析したいこともある。そこで、筆者らは、教師自らが集めた学習者の作文や言い誤りを含む会話文などのデータもコーパスとして分析できるツールの開発を進めてきた。本発表では、今回、新たに加えたタグ検索機能について紹介するとともに、運用実験の結果を報告する。

２．コーパス分析システムの概要

テキスト分析を適切に行うには、基礎データ、正確な形態素解析を可能にするタグ付け、形態素解析、の３つを統合する必要がある。発表者らが開発したコーパス分析システム「Co-Chu」（以下、Co-Chu）は、日本語テキスト分析のためのウェブアプリケーションである。Co-Chu は【Build】【Import】【Edit】【Analyze】の４機能を一つのインター

フェイスで使えるようにデザインされている（山本他 2018）。利用に際しては、まず、【Build】機能で作成するコーパスの階層構造を決める。次に、【Import】機能でデータをシステムにアップロードする。さらに、【Edit】機能を用いて、MeCab-UniDic による形態素解析を行う。解析結果に問題があれば適宜修正を行い、話し言葉や誤用を含むデータでも形態素解析が適切に行えるように編集する。その後、【Analyze】機能の活用によって、キーワード検索、コロケーション分析、複数データの比較等の分析が可能になる。

３．形態素解析の誤解析への対応

日本語学習者の発話や作文などに限らず、日本語母語話者の話し言葉にも、言い間違い、言い直し、言い淀み、フィラーなど、不規則なものが多く含まれる。そのため、これらを形態素解析した際に、誤解析が起きることがある。Co-Chuでは【Edit】機能を活用して、以下の２つの方法でこれらに対応することが可能である。タグ付けの詳細については、山本他（2018）で述べたので、以下に概略を示す。

3.1 タグ付けによる対応

収集したデータに誤用、言い直し、方言、若者ことば等が含まれており、それによって誤解析が起きた場合、Co-Chuではデータにタグ付けをし、正しい語形を示すことで、適切な形態素解析ができるようになる。タグ付けは次のような形で行う。

｜正しい語形｜（種別　実際に用いられた語形）

具体例は、次のとおりである。

（例１）ちょっとやでした。→ 解析結果：「や」は「や（助詞）」

タグ付け：｜いや｜（話　や）でした →タグ付け後の解析結果：「いや（形状詞）」

タグの種類は、研究の目的に応じて自由に設定できるが、今回は、MeCabの誤解析に対して、以下の６種類のタグを用いて分析することにした。

① 話し言葉のため誤解析が生じている場合のタグ：「話」
② 方言のため誤解析が生じている場合のタグ：「方」
③ 誤用を示すタグ：「誤」
④ 言い淀みのため誤解析が生じている場合のタグ：「淀」
⑤ 上記以外の事情で誤解析が生じている場合のタグ：「他」
⑥ 形態素解析の単語情報に誤りがあることを示すタグ：「単」

このタグ付けでは、次のような形で「タグメモ」をつけることも可能である。

｜正しい語形｜（種別　実際に用いられた語形：タグメモ）

（例２）それ忘れなく、やっぱり次の子とか

タグ付け：それ忘れ｜ないで｜（誤　なく：活用）、やっぱり次の子とか

上記の例では、「忘れないで」の活用が誤っていたので、「タグメモ」に「活用」と記載した。このタグメモによって、どんな誤用かなどの情報をメモしておくことが可能になり、後の分析に役立てることができる。

3.2 単語情報の書き換えによる対応

3.1で示したタグ付けによって、誤解析の多くに対応が可能になり、正しい形態素解析

を行えるようになる。しかし、⑥の「単」タグについては、形態素解析の誤解析がタグ付では解消できない場合に利用する。

（例３）それから焼いて、あのサラダとか→ 解析結果：「あの」は「連体詞」
タグ付け：｜あの｜（単　あの：感動詞）サラダとか

上記の例の「あの」が「あのー」のように長音になっていれば、感動詞（フィラー）と正しく解析されるが、「あの」や「その」であれば、MeCabは「連体詞」として解析する。こうした場合、【Edit】画面で解析結果を直接編集して正しいものに修正する。この修正は形態素解析そのものの書き換えではなく、結果画面のいわば表面的な修正にとどまるため、形態素解析を再度行うと、再び誤解析が生じてしまう。しかし、次に述べる分析を適切に行うためには、この書き換え作業は不可欠である。また「単」タグをつけておくことで、解析に問題があったことが示しておける。

４.「タグ」検索機能の活用

4.1 タグ検索機能の概要

Co-Chuの【Analyze】機能では、検索条件を指定することによって様々な検索が可能である。今回、この機能に、新たに「タグ」を指定した検索及び「タグメモ」の検索が可能な仕組みを整えた。「タグ」は研究者が目的に応じて、新たなタグの設定や使用しないタグの削除が可能である。

ここでは、実験用に話し言葉（母語話者同士の雑談３種、日本語母語話者（以下NS）と日本語学習者（以下NNS）の雑談３種、計６種のコーパス。録音時間232分）の文字化データをCo-Chuに取り込み、上述の６種類のタグを付けてタグ検索機能の運用実験を行った。

なお、今回の実験では、タグは原則として、形態素解析が適切に行われなかった箇所のみに付与した。つまり「話」や「方」等は、全ての「話し言葉」や「方言」に付与したわけではなく、形態素解析において誤解析が生じた場合にのみ付与した。ただし、「誤」に関しては、助詞の誤用は誤解析であるかどうかに関わらずタグ付けし、その他の誤用は誤解析になる場合のみタグ付けすることとした。

実験用タグ付きコーパスで６つのタグを検索したところ、表１の結果が得られた。

表１　タグ検索の結果

タグ / ライン数	①話	②方	③誤	④淀	⑤他	⑥単	合計
5513	153	105	29(うち助詞14)	60	28	273	649

つまり、実験用コーパスには、誤解析が635箇所で生じていた（注：助詞の誤用を除く）が、そのうち①～⑤までの361箇所はタグの付与により、誤解析を解消することができた。また、「単」タグは、そのほとんど（273のうち252）が、「感動詞」を「連体詞」と誤解析したものであった。

4.2 タグ検索機能の活用

タグ検索機能では、タグのついた形態素が一覧表示される。これを利用することによって、例えばどのような誤用があったかを見ることも容易にできる。実験用データの誤用

タグの検索結果は、次のように表示される。

結果数: 29 / ライン数: 5513

ライン	ライン番号	サブコーパス名	出現形	タグ	話者ID	タグメモ
//家族のほう|は|（誤　には：助詞）無事ですか。	79	1_J_K_tagged	には	誤	8075	助詞
そうですね、でも、やっぱり|その|（単　その：感動詞のはず）教育|を|（誤　の：助詞）専攻している人は（はい）、|その|（単　その：感動詞のはず）先生の気持ち、よく分かるんですね（あっ、はい）、こんなに準備してくれたので、やっぱりそこからもっと感謝するんじゃないですか。	155	1_J_K_tagged	の	誤	8075	助詞
|その|（単　その：感動詞のはず）、｜島｜（誤　じま：発音）に（はい）入るに、入るのに（はい）、まず、あのー、海兵隊｜に｜（誤　の：助詞）入隊しないといかんですよ〔笑〕。	257	1_J_K_tagged	じま	誤	8075	発音
|その|（単　その：感動詞のはず）、｜島｜（誤　じま：発音）に（はい）入るに、入るのに（はい）、まず、あのー、海兵隊｜に｜（誤　の：助詞）入隊しないといかんですよ〔笑〕。	257	1_J_K_tagged	の	誤	8075	助詞
いや、|その|（単　その：感動詞のはず）訓練はしないんですけど、ただ、|その|（単　その：感動詞のはず）何（なん）か｜そこ｜（誤　あそこ：指示詞）で（はい）訓練とかしたりして、で、もう見えるところに（へえー）、一回だけ入って、うん。	262	1_J_K_tagged	あそこ	誤	8075	指示詞
うん、で、その確かに、お寺か神社かどっちか覚えてないですけど（はい）、その|真ん中|（淀　まん）、その、あのー、何て言う↑、歩道↑（はい）、が、あって、その真ん中|を|（誤　に：助詞）歩けない//、は、どちらでしたっけ↑	305	1_J_K_tagged	に	誤	8075	助詞
まずどういう、まあ、私の、|あの|（単　あの：感動詞のはず）（はい）、理解の限り｜では｜（誤　に：助詞）、えー、八紘ー、あっ、これですね。	370	1_J_K_tagged	に	誤	8075	助詞
僕もやろうかなと、//そこ｜で｜（誤　に：助詞）。	305	5_N_S_tagged	に	誤	8087	助詞

図１　誤用タグの検索結果（一部）

実験用コーパスには図１に示したように助詞の誤用を中心に、発音・指示詞・単語・活用の誤りがあった。タグメモに記載された誤用の詳細をまとめると、次のとおりであった。

表２　誤用タグの内訳

誤用総数	助詞	発音	指示詞	語彙	活用
29	14	12	1	1	1

このように、タグ検索機能とタグメモの活用によって、各タグの一覧とその詳細が表示されるため、話し言葉、方言、誤用等について、どのような傾向があるかを容易に把握することができる。また、例えば、全ての誤用にタグをつけたり、タグメモを活用したりすることによって、さらに詳細な情報を整理し、分析することも可能になる。

５.おわりに

Co-Chuでは、話し言葉や誤用を含むテキストであっても、タグ付けを行うことによって形態素解析を正確に行うことが可能になり、適切な言語分析をすることができる。今後さらにどのような活用方法が可能であるか、研究を継続していきたい。

謝辞　本研究の一部は、科学研究費基盤研究（C）18K00723の助成を受けている。

引用文献・引用URL

山本裕子・川村よし子・小森早江子・本間妙(2018)「話し言葉や誤用の含まれたテキストに対応可能なコーパス分析システムの開発」『2018年度日本語教育学会秋季大会予稿集』pp.295-300.日本語教育学会.

MeCab　http://taku910.github.io/mecab/　（2019 年 3 月 31 日閲覧）
UniDic　https://unidic.ninjal.ac.jp（2019 年 3 月 31 日閲覧）